DOCUMENT RESUME

ED 442 717 SO 031 874

AUTHOR	Ramey, Madelaine
TITLE	Reducing the White-Nonwhite Achievement Gap.
PUB DATE	2000-04-00
NOTE	21p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Black Students; Correlation; Educational Research; Grade 4; Intermediate Grades; Public Schools; Test Validity; *White Students
IDENTIFIERS	Dependent Variables; Seattle Public Schools WA

ABSTRACT

It is well documented that there continues to be a gap between white and nonwhite student achievement. A study develops and tests a measure of white-nonwhite achievement gap reduction. The ultimate purpose is to use the measure as the dependent variable in a qualitative study of what works in reducing the gap. The strategy used in addressing this purpose is to propose a measure of gap reduction, examine its properties, use it to identify classrooms that appear successful in reducing the gap, and to study these classrooms in an attempt to identify key characteristics. Data were supplied by the Seattle School District (Washington); their 1999-2000 enrollment was approximately 47,000 students, of which 23% identified themselves as black. The study focused on students who were in the fourth grade in 1998-1999. The measure of gap reduction appears adequate to the task of identifying classrooms that narrow the test score gap between children of color and white children. That part of the variance in the measure that is attributable to classrooms is considerable. Gap reduction depends mostly on classroom factors, as distinct from characteristics of the student. Moreover, the composite classroom gap reduction index correlates highly with a measure of overall classroom achievement growth. Correlation size suggests that success in reducing the gap tends to occur with success in increasing achievement overall, but the two do not always coincide. This has implications for the continuation of the study, which will use a control group of classrooms with high overall achievement gains and not so high gap reduction indices. The plan is to add student mobility rate to the set of classroom context variables as soon as it is available. Stability of the gap reduction measure over time needs to be determined. (Contains 20 references.) (BT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 442 717

REDUCING THE WHITE-NONWHITE ACHIEVEMENT GAP

Madelaine Ramey

Rain City Associates
13545 Corliss Avenue N.
Seattle, WA 98133

206-364-0262

SO 031 874

Madelaine Ramey

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the
American Educational Research Association, New Orleans, 2000.

BEST COPY AVAILABLE

REDUCING THE WHITE-NONWHITE ACHIEVEMENT GAP

Introduction

That there continues to be a gap between white and nonwhite student achievement is well documented (Campbell, Voelkl, & Donahue, 1997; Nettles, 1988). There has been some research regarding what works in reducing the gap (Ferguson, 1998; Finn & Achilles, 1990; Grissmer, Flanagan, and Williamson,1998; Ladson-Billings, 1994; Ramey, 1992), although approaches vary along several dimensions--one of which is in how gap and gap reduction are measured or defined.

Although the immediate purpose of this study is to develop and test a measure of white-nonwhite achievement gap reduction, the ultimate purpose is to use the measure as the dependent variable in a qualitative study of what works in reducing the white-nonwhite achievement gap. The strategy used in addressing this ultimate purpose is to propose a measure of gap reduction, examine its properties, use it to identify classrooms that appear successful in reducing the gap, and to study these classrooms in an attempt to identify key characteristics. Specifically, the study addresses the following questions:

1. What is an appropriate measure of ethnic achievement gap reduction?

2. What portions of the variance in the gap reduction measure are associated with students and classrooms? How much of the variance attributable to classrooms is due to factors under teacher or school control? How reliable are indices of classroom effectiveness in reducing the gap?

3. To what extent do gap reduction indices correlate with indices of overall (across ethnic groups) classroom effectiveness in promoting student achievement growth?

An additional question, to be addressed during summer 2000, is,

4. How consistent are gap reduction indices from one year to the next?

Method

Data Sources

Data were supplied by the Seattle School District. The District's 1999-2000 enrollment was approximately 47,000 students, 60% of whom identified themselves as other than White--24% Asian, 23% Black, 10% Latino/Chicano, and 3% Native American (Seattle School District, 1999) . The ethnic distribution for certificated teachers was 78.6% White and 21.4% minority—8% Asian, 9.6% Black, 2.6%

Latino/Chicano, and 1.1% Native American (S. Fong, personal communication, January 21, 2000).

The study focused on students who were in the 4th grade in 1998-1999. All 4th grade students are required to take the Washington Assessment of Student Learning (WASL), a criterion-referenced test containing multiple choice, short answer, and extended response items. The multiple choice item responses are machine-scored; the short answers and extended responses are hand-scored by state-trained scorers. KR-20 reliability coefficients for reading and math exceed .80 (State of Washington, Superintendent of Public Instruction, 1997) Most of the district's fourth grade students were administered the Iowa Tests of Basic Skills (ITBS) the two previous years, in the second and third grades. In all, scores for both reading and math from the 1999 administration of the WASL (Grade 4) and the 1997 and 1998 administrations of the ITBS (Grades 2 and 3, respectively) were used in this study.

Analyses

Each of the questions listed in the Introduction called for a specific analysis procedure.

What is an appropriate measure of gap reduction?

The gap reduction measure must be a valid representation of change, at the individual child level, in white-nonwhite achievement differences. This requirement actually embodies three criteria. First, the measure must be one of gain rather than status at a single point in time because a given classroom can be held more accountable for a change that occurred while a child was in that classroom (Duncan & Raudenbush, 1999). Second, it must attach to an individual child, in order to ultimately isolate the contributions of child background and classroom (Bryk & Raudenbush, 1992). Third, it must incorporate a comparison of a nonwhite child's achievement with some standard of white student achievement.

These criteria led to the decision to use a standardized residual gain measure, obtained by regressing spring 1999 test scores of white students who had been in the same classroom throughout school year 1998-1999 on their spring 1998 test scores, and representing a nonwhite student's score as a standardized (divided by standard error) deviation from that predicted for a white child with the same 1998 score. (Although the figures are shown as part of the Results section, the reader might wish to refer to Figures 1 and 2, Appendix A, to see a graphical representation of a gap score.)

The primary reasons for using this kind of residual rather than a simple gain measure were two: Simple gains are negatively correlated with initial scores and the available data are such that the test taken in 1999 is not the same as that taken in 1998.

Gap **reduction** required measurements of gap for two years, 1998-99 and 1997-98. Gap scores for 1997-1998 were computed as for 1998-1999 and on the same students for whom 1998-1999 gap scores were available. The major difference was that computation of 1997-1998 gap scores used the ITBS in both 1997 (end of 2nd grade) and 1998 (end of 3rd grade).

What portions of the variance in the gap reduction measure are associated with students and classrooms? How much of the variance attributable to classrooms is due to factors under teacher or school control? How reliable are indices of classroom effectiveness in reducing the gap?

A partitioning of variance was of interest because, if we were going to assess effects of classroom in reducing the gap, we wanted to be sure that some of the variance in the gap reduction measure was indeed due to this source. Furthermore, even if a significant portion of the variance was attributable to classrooms (teachers), we still wanted some assurance that the classroom variance was due to alterable factors. Finally, if we were to use indices of classroom effectiveness in reducing the gap to identify classrooms that are exemplary, we wanted to be sure that these indices were reliable.

An underlying model. A number of writers recommend a multilevel approach for isolating the contributions of students and classrooms to achievement outcomes (Bryk and Raudenbush, 1992; Goldstein and McDonald, 1988; and Aitkin and Longford, 1986). The approach is called multilevel, or hierarchical, because it can deal with data sets in which observations at one level (student data) are subsumed under groups (classrooms or schools) at another level.

To this end, the gap reduction measure described in the preceding section was used as dependent variable in a three-level hierarchical linear modeling (HLM) analysis, which treats time (year) as "nested" within students which are in turn nested within classrooms (Bryk & Raudenbush, 1992, p. 185).

$$Y = \pi_0 + \pi_1 (YEAR) + e, \qquad (1)$$

where

Y is a gap score for reading or math in a given year, 1997-1998 or 1998-1999, for a particular nonwhite child in a particular classroom. (A gap score for 1997-1998 was based on two administrations of the ITBS, one in 1997 and one in 1998; a gap score for 1998-1999 was based on the ITBS in 1998 and the WASL in 1999.). Scores for each school year were standardized as noted on p. 2.

(YEAR) takes on the value 0 in 1998 and 1 in 1999;

π_0 is the expected initial status of the child, i.e., the expected 1998 ITBS gap score for the child;

π_1 is the expected change in gap score (gap reduction) from 1998 to 1999 ; and

e is the random effect that represents the deviation of the child's gap score from that expected.

In order to isolate the contribution of the individual child (as distinct from classroom or school) to this change in gap score from 1998 to 1999, we assumed that π_1 could be represented as a function of measured characteristics of the child plus a random error component; that is,

$$\pi_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + r, \qquad (2)$$

where β's are regression coefficients, X's are the measured characteristics of the child, k is the number of measured characteristics, and r is a random effect that represents the deviation of a student's score from that predicted.

(A similar equation could be given for π_0, but initial standing is not of primary interest in this application.)

The following student background variables were considered as candidates for entry as Xs in equation (2).

- Gender (1 if male, 0 if female) (GENDER)
- Family Status (1 if living with both parents, 0 otherwise) (BOTH)
- Free or reduced price lunch status (1 if eligible, 0 otherwise) (FRL)
- Limited English Proficient (LEP) status (1 if LEP, 0 otherwise) (LEP)
- Special Education status (1 if Special Education, 0 otherwise) (SPED)
- Continuing student (1 if attended same school previous year, 0 otherwise) (CONT)

In order to isolate the classroom contribution to the change in gap score from 1998 to 1999, we assumed that this contribution operates through the intercept, β_0, in equation (2). (See Burstein,1980, for a compelling rationale for this assumption.) That is,

$$\beta_0 = \gamma_0 + \gamma_1 W_1 + \gamma_2 W_2 + \ldots + \gamma_m W_m + u, \qquad (3)$$

where the γs are regression coefficients, W's are school context variables, m is the number of school level predictors, and u is a random effect—that part of the classroom level intercept, β_0, not predicted from the school context variables. (Ultimately, classroom effectiveness indexes are estimates of u.)

Seven classroom context variables were considered. All are aggregated student background variables.

- Percentage male students (PCTMALE)
- Percentage students living with both parents (PCTBOTH)
- Percentage students eligible for free or reduced price lunch (PCTFRL)
- Percentage minority students (PCTMIN)
- Percentage LEP students (PCTLEP)
- Percentage Special Education students (PCTSPED)
- Percentage continuing students (PCTCONT)

Student mobility rate has also been shown a relevant predictor of achievement at the classroom and school levels (Ramey, 1998; Ingersoll, Scammon, & Eckerling, 1989). It was not included here because it has not yet been calculated for 1999. Mobility rate will be examined as a classroom-level predictor of gap reduction in a continuation of the study.

As can be seen, equations 1 through 3 capture the hierarchical nature of the analysis. Also, when equation 3 is substituted into equation 2 and the resulting equation 2 is substituted into equation 1, one has a mixed model equation for Y. It might be noted that a mixed model equation is the basis for the Tennessee Value-Added Assessment System, which purports to provide unbiased estimates of the effects of schools and teachers on the academic growth of students and to distinguish between these effects and those of outside influences (McLean, Sanders, & Stroup, 1991).

The above-listed variables—the X's (student level) and W's (classroom level) -- were excluded as predictors at their respective levels if they proved nonsignificant through exploratory HLM runs. The variables that survived as predictors are listed in Results.

Determine relative contributions to variance of gap reduction measure. Because of its hierarchical nature, HLM can partition gap score variance into individual student and classroom components. Two simple (predictors excluded) runs of HLM were performed to provide such a partitioning for reading and math. Two full model runs of HLM provided contributions of the predictors to these components of variance.

Estimate classroom effects and calculate reliabilities of the estimates. In addition to a partition of variance, other results of the HLM analysis are the quantities needed to calculate empirical Bayes estimates of classroom effects on gap reduction and the reliabilities of these estimates. In essence, the estimate of a classroom's effectiveness is what is "left over" in the classroom's average change in gap score after the effects of unalterable compositional variables (e.g., percent of students eligible for free or reduced-price lunch) are removed. This estimate is multiplied by λ_k, its reliability. Specifically, an empirical Bayes estimate of classroom k's effect on gap reduction is

$$u_k^* = \lambda_k \, \text{est.}(u_k)$$

where

$$\lambda_k = \text{est. var}(u)/\{\text{est. var(u)} + [\text{est. var}(r) + \text{est. var}(e)]/n_k\},$$

$$\text{est}(u_k) = \text{observed average gap score change} - \text{est. average gap score change},$$

and *est.* = estimated, n = number of gap reduction scores (nonwhite students) in the classroom, and all estimated quantities are provided by the HLM analysis. (See Bryk & Raudenbush, 1992, pp. 80, 125, & 178.) **The classroom effect estimates , u_k^*, are what we will refer to as classroom gap reduction indices.**

As can be seen, classroom gap reduction indices are obtained by subtraction; the estimates consist of what is "left over" in the classroom's average observed value after the effects of unalterable composition variables are removed. The rationale is that, since the effects of unalterable factors have been removed, that which is left over is due to alterable factors such as classroom policies and practices. Three assumptions underlie this rationale.

The first assumption is that all relevant student background and classroom composition variables have been included. For this reason, it is important to add student mobility rate as a classroom composition variable as soon as possible.

The second assumption is that there is no interaction between student background and classroom policies and practices. Pituch (1999) showed that an interaction between student background and classroom practice exists when (a) within-classroom slopes vary across classrooms and (b) the variation in these slopes remains after adding classroom composition variables. Accordingly, tests for such interactions were conducted before computing the final estimates of classroom effectiveness.

The third underlying assumption is that there is no interaction between classroom composition and classroom policies and practices. Raudenbush and Willms (1995) showed that if this assumption is false, the estimator is biased. It remains to be determined, when relevant policies and practices are identified, the degree to which they are correlated with composition variables. This determination will be possible in the continuation of the study.

The gap reduction indices were evaluated in terms of their reliabilities. Relatively low reliabilities for reading and math suggested use of a composite index, the sum of the indices for reading and math. The reliability of such a composite is given by Mosier (1943). The size of the Mosier reliability coefficient was used to evaluate the composite indexes.

To what extent do gap reduction indices correlate with indices of overall classroom effectiveness?

The answer to this question required of course that indices of overall classroom effectiveness be computed. Computations were based on the same basic model and followed the same analysis steps as for the gap reduction indices. That is, equations (1) through (3) were used, but the dependent variable, Y, and definitions of π_0 and π_1 changed.

- Y became a test score for reading or math in a given year, 1998 or 1999, for a particular child, white or nonwhite, in a particular classroom. (A test score for 1998 was based on the ITBS; a test score for 1999 was based on the WASL.). Scores for each year were standardized using their own means and standard deviations.

- π_0 became the expected 1998 ITBS score for the child.

- π_1, the parameter of interest, was the expected change in test score from 1998 to 1999.

- The same individual child and classroom level variables, X and W, were considered as possible predictors in the new equations 2 and 3.

Use of a hierarchical model permitted a partitioning of the variance in test score gains into individual student and classroom components. Use of HLM as an analysis procedure permitted calculation of empirical Bayes estimates of classroom effects—indices of classroom effectiveness In promoting achievement gain—and reliabilities of these estimates.

Having computed the estimates, it remained only to correlate them with the classroom gap reduction indices. The correlation was obtained using SPSS's Pearson R procedure.

How consistent are gap reduction indices from one year to the next?

This question concerns another kind of reliability. If the teacher is the same from one year to the next, we would expect that if the teacher is the primary agent of gap reduction at the classroom level, then rank of the classroom index should remain approximately the same. Credibility of classroom selections based on the indices depend on there being some degree of year-to-year stability in the indices.

Classroom gap reduction indices for the following year, 1999-2000, will be computed as described for 1998-1999. A subset of classrooms will be compared with respect to the ranks of their indices in the two years. The subset of classrooms will be those with the same teacher for both years, in schools with the same principal and a relatively low level of staff mobility from 1999 to 2000.

Results

Results are presented below by the question addressed.

What is an appropriate measure of gap reduction?

As noted in the Methods section, the criteria set for an appropriate measure of gap reduction led to the decision to use a standardized residual gain measure, obtained by regressing spring 1999 test scores of white students who had been in the same classroom throughout school year 1998-1999 on their spring 1998 test scores, and representing a nonwhite student's score as a deviation from that predicted for a white child with the same 1998 score. Figures 1 and 2 show the scatterplots for reading for white and nonwhite students, respectively, with the best-fit line generated for white students drawn in both scatterplots. The unstandardized gap score for one child (with a 1998 score of 13) is shown in Figure 2 as the distance between the point representing his score and the best-fit line for white students. (See end of paper for Figures 1 and 2.)

What portions of the variance in the gap reduction measure are associated with students and classrooms? How much of the variance attributable to classrooms is due to factors under teacher or school control? How reliable are indices of classroom effectiveness in reducing the gap?

Predictors for the model. Through exploratory HLM runs, three student background variables LEP, SPED, and FRL, were found predictive of *individual student level* gap score change in reading. Only one background variable, LEP, was found predictive in math. According to the tests suggested by Pituch (1999), none of these variables appeared to interact with school policies and practices.

Thus, for reading, equation 2 became

$$\pi_1 = \beta_0 + \beta_1(\text{LEP}) + \beta_2(\text{SPED}) + \beta_3(\text{FRL}) + r,$$

and for math,

$$\pi_1 = \beta_0 + \beta_1(\text{LEP}) + r,$$

Two *classroom level* predictors, PCTFRL and PCTCONT, were significant for reading; for math, three were significant: PCTBOTH, PCTFRL, and PCTLEP

Thus, for reading, equation 3 became

$$\beta_0 = \gamma_0 + \gamma_1(\text{PCTFRL}) + \gamma_2(\text{PCTCONT}) + u$$

and for math,

$$\beta_0 = \gamma_0 + \gamma_1(\text{PCTBOTH}) + \gamma_2(\text{PCTFRL}) + \gamma_3(\text{PCTLEP}) + u.$$

Thus change in gap score, π_1 for reading is a function of three individual child characteristics, LEP, SPED, and FRL, as well as two classroom composition effects, PCTFRL and PCTCONT. Change in gap score for math is a function of one individual child characteristic, LEP, and three classroom composition effects, PCTBOTH, PCTFRL, and PCTLEP. The importance of English proficiency and the socioeconomic variable, FRL, at both the individual child and classroom levels is worthy of note.

Relative contributions to variance. For reading, an unconditional HLM analysis indicated that the percentage of variation in gap reduction attributable to classrooms is 52%. The remaining 48% is due to students within classrooms.

Of the between-classroom variation, 36% is due to classroom composition, represented by percentage of students eligible for free or reduced-price lunch and percentage of students who are continuing in the same school. This means that up to 64% (100%-36%) of the between-classroom variation (or 64 x 52 = 33% of the total variation) could be due to factors under school or teacher control.

For math, a similar HLM analysis indicated that 59% of the variance in math gap reduction is attributable to classrooms. The remaining 41% is due to students within classrooms.

Of the between-classroom variance, 9% is due to classroom composition, represented by percentage of students from 2-parent families, percentage eligible for free or reduced-price lunch, and percentage limited English proficient. This means that up to 91% (100%-9%) of the between-classroom variation (or 91 x 59 =54% of the total variation) could be due to school programs and practices--factors under school or teacher control.

Table 1 shows, for both reading and math, the percentage of total variation in the gap reduction measure that can be attributed to the three sources, students, classroom composition, and other classroom characteristics (programs and practices). It seems clear from Table 1 that there is enough possible variation due to alterable classroom characteristics to justify computing classroom gap reduction indices.

Classroom gap reduction indices and their reliabilities. For reading, 187 classrooms had gap reduction indices ranging from –1.277 to 1.187, with a mean of -.025 and standard deviation of .366. The reliabilities of these estimates ranged from .439 to .829, with a mean of .67.

For math, 183 classrooms had gap reduction indices ranging from –1.361 to 1.179, with a mean of .011 and standard deviation of .489. The reliabilities of these estimates ranged from .585 to .897, with a mean of .78.

Table 1

Percentage of Total Variation in Reading and Math Gap Reduction Scores Due to Students, Classroom Composition and Other Classroom Characteristics

Test	Students within Classrooms	Classrooms	
		Composition	Other Characteristics
Reading	48	52	
		19	33
Math	41	59	
		5	54

The reliabilities for math are higher than the reliabilities for reading. Using the conventional .80 as criterion for acceptable reliability, classroom gap reduction indices do not have acceptable reliability for reading, although they come close for math.

To boost the reliabilities, the indices for reading and math were summed to produce a composite index of gap reduction. Reliabilities of the sums were calculated using Mosier's (1943) formula. The 183 classrooms with both a reading and a math gap reduction index had composite gap reduction indices ranging from –2.05 to 1.732, with a mean of -.006 and standard deviation of .71. The reliabilities of these estimates ranged from .656 to .906, with a mean of .81.

Since the mean reliability of the combined indices exceeds the usual .8 criterion, some tentative statements about the relative effectiveness of the 183 classrooms in reducing gap might be made. For example, applying the one standard deviation criterion to the composite indices, it could be concluded that 27 classrooms (the number whose composite index exceeds one standard deviation) are doing better than expected, given their composition, in reducing the white-nonwhite achievement gap.

To what extent do gap reduction indices correlate with indices of overall classroom effectiveness?

Predictor variables. As noted in the Methods section, computation of overall classroom effectiveness indices was based on the same basic model and followed the same procedural steps as for the gap reduction indices. What changed were the dependent variable, Y, and the definitions of π_0 and π_1. Also, the analysis provided for a different subset of { X }, the individual child level predictors, and { W }, the classroom level predictors. As with the gap reduction measure, the tests suggested by Pituch

(1999) produced results indicating no interaction between the child level variables and school policies and practices.

In particular, π_1, the expected change in reading test score from 1998 to 1999 was found to be

$$\pi_1 = \beta_0 + \beta_1(\text{LEP}) + \beta_2(\text{SPED}) + \beta_3(\text{FRL}) + \beta_4(\text{BOTH}) + r,$$

and for math,

$$\pi_1 = \beta_0 + \beta_1(\text{GENDER}) + r,$$

One classroom level predictor, PCTBOTH, was significant for reading; for math, three were significant: PCTBOTH, PCTFRL, and PCTLEP

Thus, for reading, equation 3 became

$$\beta_0 = \gamma_0 + \gamma_1(\text{PCTBOTH}) + u$$

and for math,

$$\beta_0 = \gamma_0 + \gamma_1(\text{PCTBOTH}) + \gamma_2(\text{PCTFRL}) + \gamma_3(\text{PCTLEP}) + u.$$

Relative contributions to variance. For reading, the unconditional HLM analysis indicated that the percentage of variation in test score change attributable to classrooms is 88%. The remaining 12% is due to students within classrooms.

Of this between-classroom variation, 21% is due to classroom composition, represented by percentage of students with limited English proficiency. This means that up to 79% (100%-21%) of the between-classroom variation (or 79 x 88 = 70% of the total variation) could be due to school programs and practices.

For math, A similar HLM analysis indicated that 90% of the variance in math test score change is attributable to classrooms. The remaining 10% is due to students within classrooms.

Of this between-classroom variance, 19% is due to classroom composition, represented by percentage of students from 2-parent families, percentage eligible for free or reduced-price lunch, and percentage limited English proficient. This means that up to 81% (100%-19%) of the between-classroom variation (or 81 x 90 =73% of the total variation) could be due to school programs and practices.

It seems clear that there is enough possible variation due to alterable classroom characteristics to justify computing overall classroom effectiveness indices.

Classroom effectiveness indices and their reliabilities. For reading, 190 classrooms had overall effectiveness indices ranging from –.761 to .943, with a mean of .051 and standard deviation of .213. The reliabilities of these estimates ranged from .46 to .84, with a mean of .69.

For math, 187 classrooms had overall effectiveness indices ranging from –.829 to .652, with a mean of -.042 and standard deviation of .286. The reliabilities of these estimates ranged from .58 to .90, with a mean of .78.

As with the gap reduction indices, the reliabilities for math are higher than those for reading. Using the conventional .80 as criterion for acceptable reliability, overall classroom effectiveness indices do not have acceptable reliability for reading, although they come close for math.

To increase the reliabilities, the indices for reading and math were summed to produce a composite index of overall classroom effectiveness, and Mosier's (1943) formula was used to calculate reliabilities of the composites. The 187 classrooms with both a reading and a math index had composite effectiveness indices ranging from –1.59 to 1.378, with a mean of .01 and standard deviation of .43. The reliabilities of these estimates ranged from .67 to .91, with a mean of .82.

Correlation of the two indices. The Pearson product moment coefficient of correlation between the two indices is .84. Of the 27 classrooms whose composite gap reduction index exceeds one standard deviation, 18 have composite overall classroom effectiveness indices that exceed one standard deviation.

Summary and Conclusions

The measure of gap reduction, developed to satisfy the three criteria described earlier, appears adequate to the task of identifying classrooms that narrow the test score gap between children of color and white children. That part of the variance in the measure that is attributable to classrooms is considerable; i.e., gap reduction depends in large part on classroom factors, as distinct from characteristics of the student.

Moreover, the composite classroom gap reduction index correlates highly, .84, with a measure of overall classroom achievement growth. The size of the correlation suggests that success in reducing the gap tends to occur with success in increasing achievement overall, but the two do not always coincide. This has important implications for the continuation of the study, which will use a control group of classrooms with high overall achievement gains and not so high gap reduction indices.

However, as noted earlier, the rationale underlying computation of gap reduction indices assumes that all relevant student background and classroom context variables have been included and that there is no interaction between classroom composition and classroom policies and practices. Accordingly, we plan to add student mobility rate to the set of classroom context variables as soon it is available; and, since the ultimate aim

of the larger study is to identify relevant policies and practices, it should be possible to check on the null interaction assumption. If the assumption is found untenable, we will see how the gap reduction measure is affected, and, if necessary, adjust it.

Also noted earlier, stability of the gap reduction measure over time needs to be determined. This requires test scores from an additional year, 2000. When these scores are added to the data base—sometime in summer 2000—we will be able to determine the degree to which ranks of classrooms' gap reduction scores for 1999 correlate their ranks for 2000. At the same time, we will

References

Aitkin, M. & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. Journal of the Royal Statistical Society, Series A, 149, 1 – 43.

Bryk, A., Raudenbush, & Congdon, R. (1996). HLM: Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L Programs. Chicago, IL: Scientific Sortware International.

Bryk, A. & Raudenbush, S. (1992). Hierarchical linear models: Applications and data analysis methods. Newbury Park, CA: Sage.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. Review of Research in Education, 8, 158 – 233.

Campbell, J.R., Voelkl, K.E., & Donahue, P.L. (1997). NAEP trends in academic progress. Washington, DC: National Center for Education Statistics.

Dallas Public Schools. (1997). The Dallas value-added accountability system. Division of Research, Evaluation and Information Systems.

Duncan, G.J. & Raudenbush, S.W. (Winter 1999). Assessing the effects of context in studies of child and youth development. Educational Psychologist 34(1), 29-41.

Ferguson, R.F. (1998). Teachers' perceptions and expectations and the black-white test score gap. In C. Jencks & M. Phillips (Eds.), The Black-White Test Score Gap (pp. 273-317). Washington, DC: Brookings Institution.

Finn, J.D. & Achilles, C.M. (Fall 1990). Answers and questions about class size: A statewide experiment. American Educational Research Journal 27(3), 557-577.

Goldstein, H. & McDonald, R. P. (1988). A general model for the analysis of multilevel data. Psychometrika, 53(4), 455 – 467.

Grissmer, D.,Flanagan, A., & Williamson, S. (1998). Why did the black-white score gap narrow in the 1970s and 1980s? In C. Jencks & M. Phillips (Eds.), The Black-White Test Score Gap (pp. 273-317). Washington, DC: Brookings Institution.

Ingersoll, G. M., Scammon, J. P., & Eckerling, W. D. (1989). Geographic mobility and student achievement in an urban setting. Educational Evaluation and Policy Analysis, 11(2), 143-149.

Ladson-Billings, G. (1994). The dreamkeepers: Successful teachers of African-American children. San Francisco: Jossey-Bass.

McLean, R. A., Sanders, W. L. & Stroup. W. W. (1991). A unified approach to mixed linear models. American Statistician, 34, 54 – 64.

Mosier, C. (1943). On the reliability of a weighted composite. Psychometrika, 8,(3), 161 – 168.

Pituch, K. (1999). Describing school effects with residual terms; Modeling the interaction between school practice and student background. Evaluation Review, 23(2), 190-211.

Ramey, M. C. (1998). Factors that impact on or relate to academic achievement: Part 2—Indexes of classroom and school contributions. Seattle, WA. Seattle Public Schools.

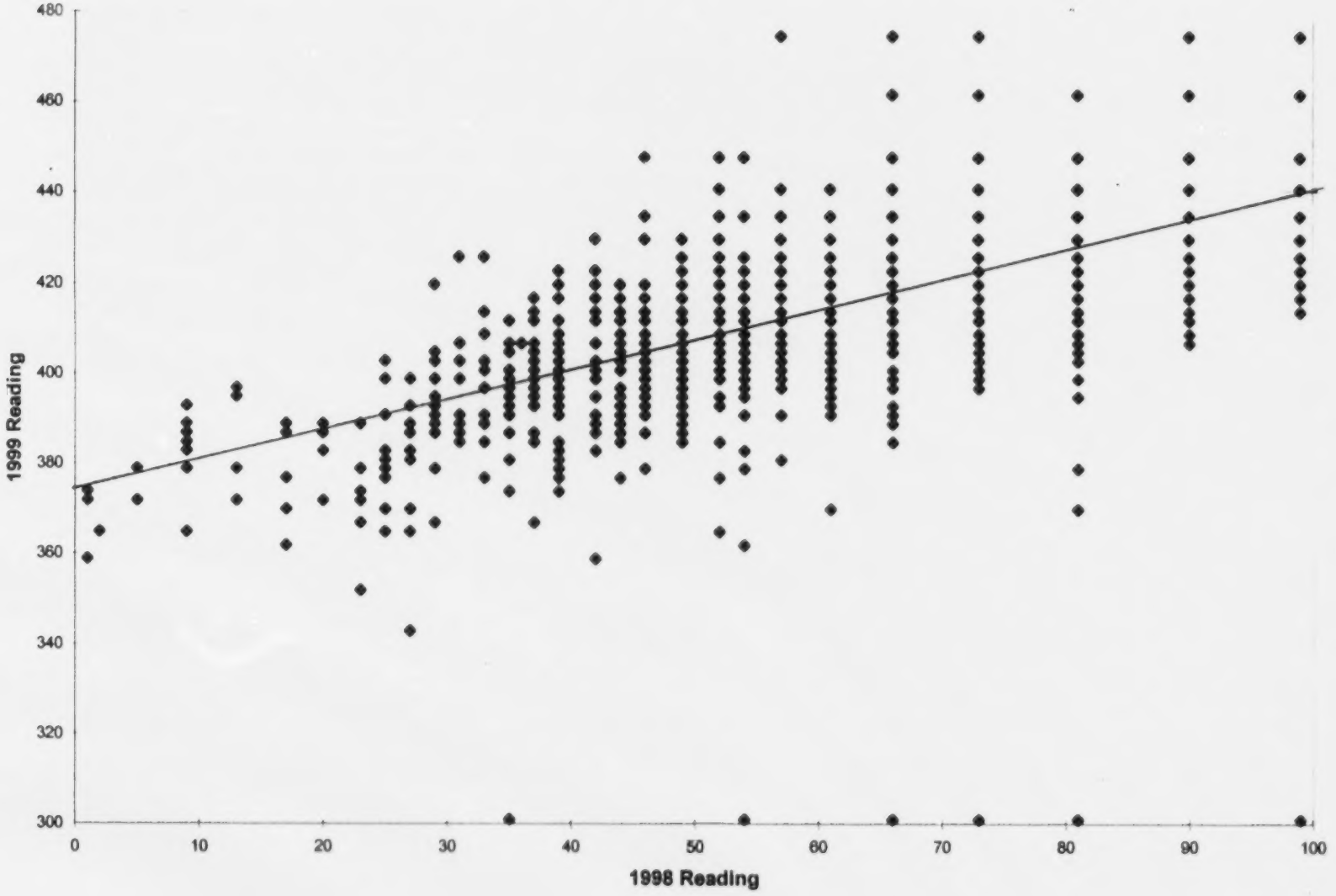
Ramey, M. (1992, April). Classroom characteristics related to ethnic achievement gap reduction. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Raudenbush, S.W. & Willms, J.D. (1995). The estimation of school effects. Journal of Educational and Behavioral Statistics, 30(4), 307-335.

State of Washington, Superintendent of Public Instruction. (1997). Washington Assessment of Student Learning. Author.

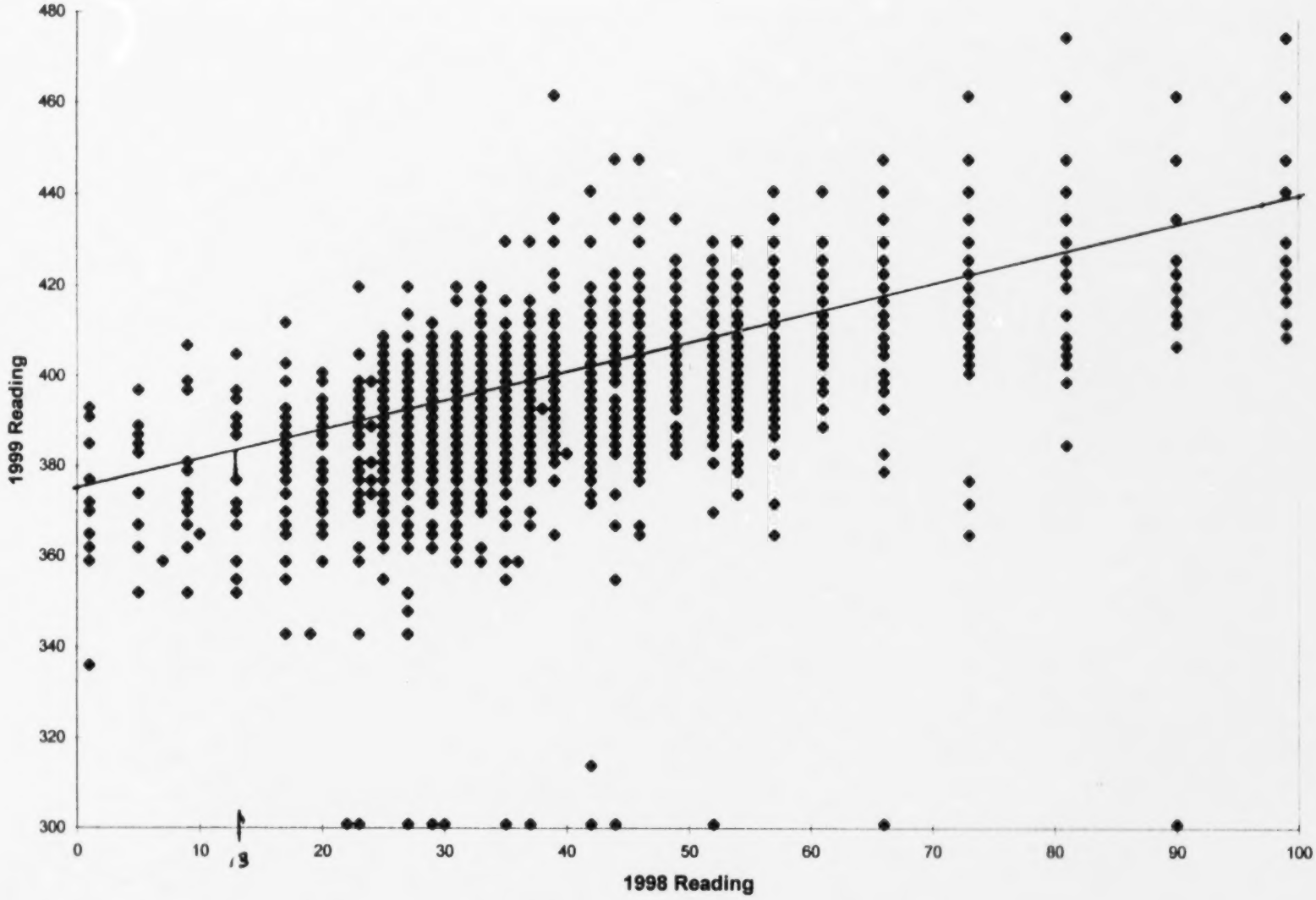
Appendix A

Figure 1. 1999 Reading Versus 1998 Reading--White Students



Equation for best fit line is Y = 374.7 + .64 X

Figure 2. 1999 Reading Versus 1998 Reading--Nonwhite Students



Equation for line is Y = 374.7 + .64 X